Efficient Bulk Data Replication for the Earth System Grid

Alex Sim, Dan Gunter, Vijaya Natarajan, Arie Shoshani,
Dean Williams, Jeff Long, Jason Hick, Jason Lee, Eli Dart

Alex Sim

Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

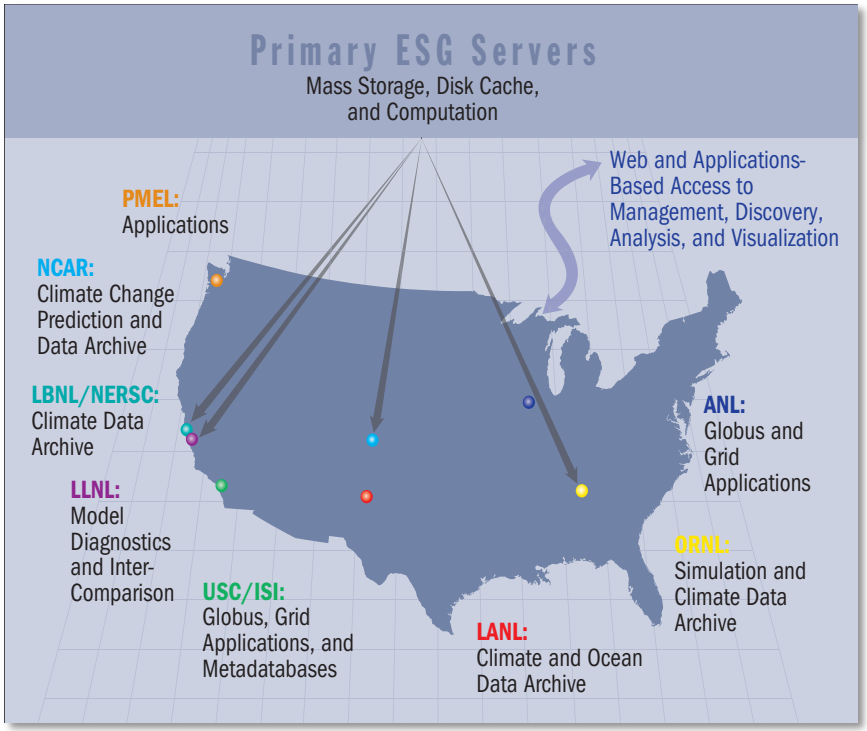
Outline

- Earth System Grid
- Bulk Data Movement
- Monitoring
- Use case
- Results

Earth System Grid

- **Earth System Grid (ESG)**
  - To support the infrastructural needs of the national and international climate community, ESG is providing crucial technology to securely access, monitor, catalog, transport, and distribute data in today's grid computing environment.
  - ANL, LANL, LBNL, LLNL, NCAR, ORNL, PMEL, USC/ISI
- **ESG's mission is to provide climate researchers worldwide with access to**
  - Data,
  - Information,
  - Models,
  - Analysis tools, and
  - Computational resources

  required to make sense of enormous climate simulation datasets.
- **Project history**
  - ESG-I (1999-2001)
  - ESG-II (2001-2006)
  - ESG-CET (2006-present)
- **Production since 2004**



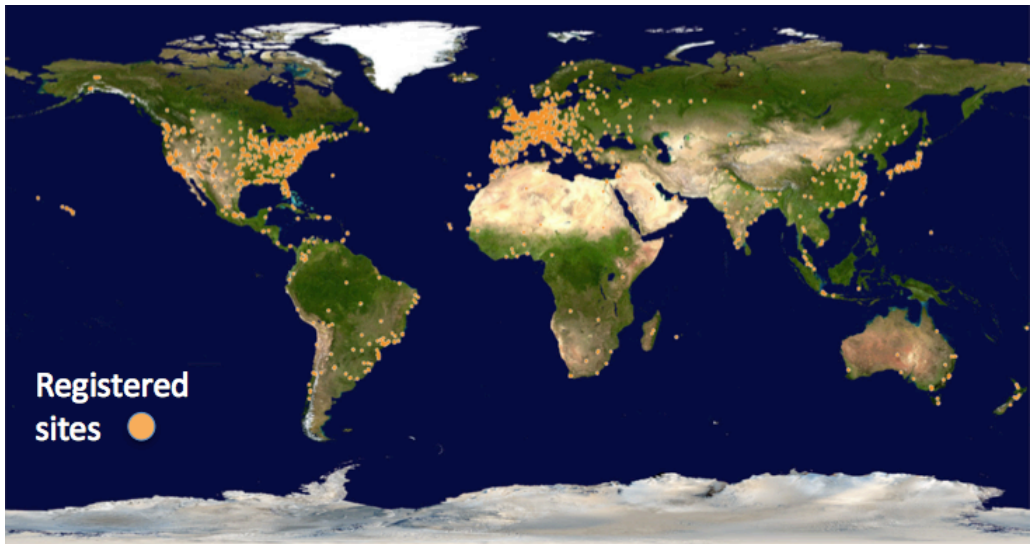
ESG current statistics

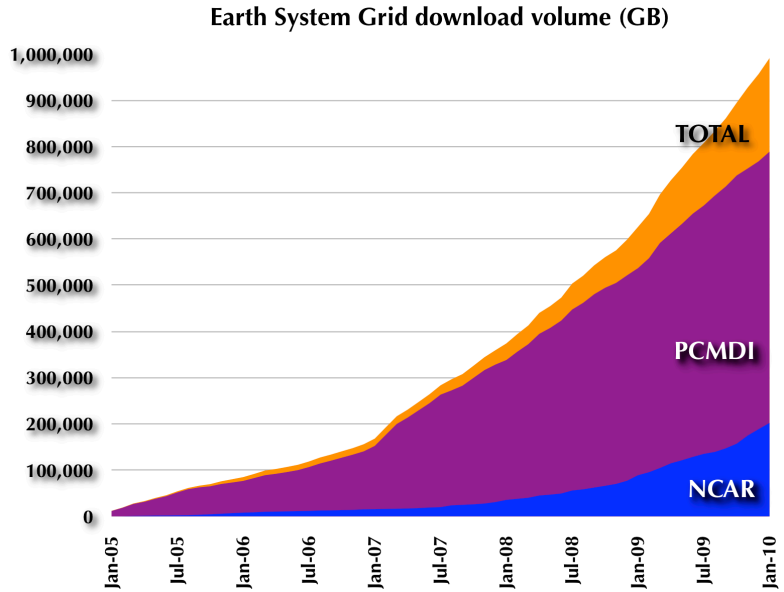
- **NCAR CCSM ESG portal**
  - 237 TB of data at four locations (NCAR, LBNL, ORNL, LANL) : 965,551 files
    - Includes the past 7 years of joint DOE/NSF climate modeling experiments
- **LLNL CMIP-3 (IPCC AR4) ESG portal**
  - 35 TB of data at one location
    - 83,337 files, model data from 13 countries
    - Generated by a modeling campaign coordinated by the Intergovernmental Panel on Climate Change (IPCC)
    - Over 545 scientific peer-review publications
- **Serving data to the community**
  - Coupled Model Intercomparison Project, Phase 3 (CMIP-3)
  - Community Climate System Model (CCSM)
  - Parallel Climate Model (PCM)
  - Parallel Ocean Program (POP)
  - The North American Regional Climate Change Assessment Program (NARCCAP)
  - Cloud Feedback Model Intercomparison Project (CFMIP)
  - Carbon-Land Model Intercomparison Project (C-LAMP)

Courtesy: Gary Strand - NCAR

ESG current download statistics

- **Geographic distribution of the users that downloaded data from ESG web portals**
  - Over 2,700 sites
  - 120 countries
  - 16,000 users
  - Over 1 PB downloaded

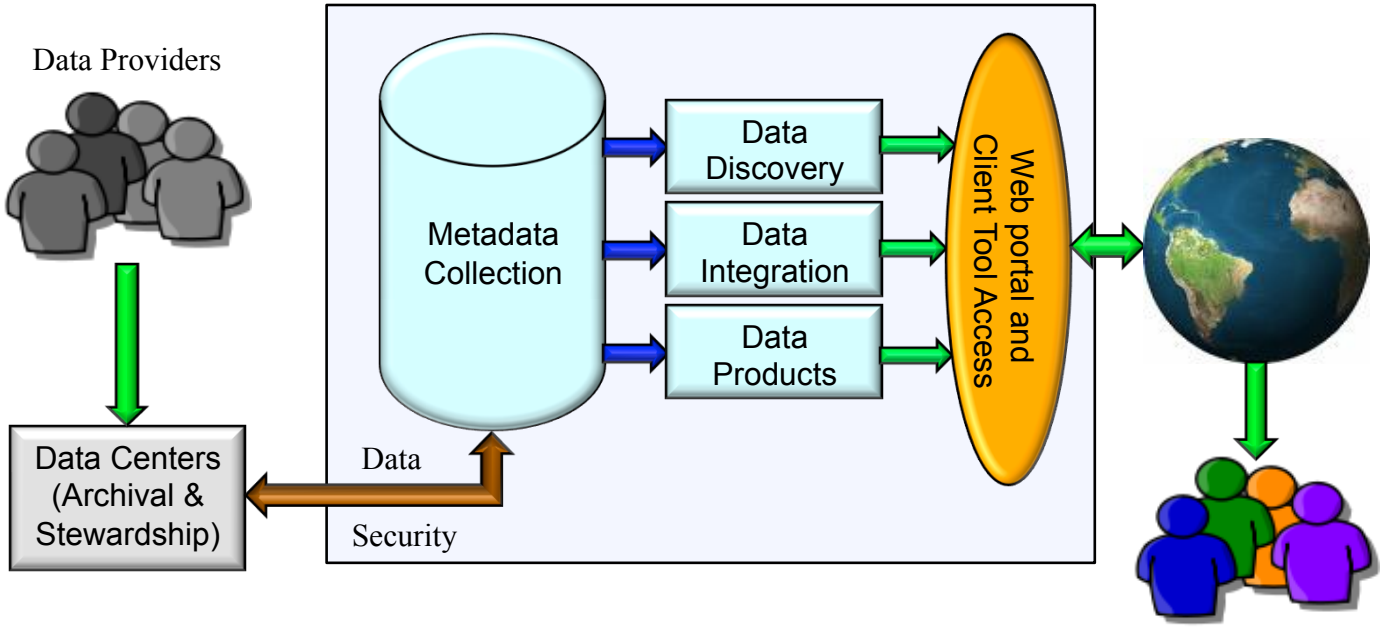




ESG Enabling Technologies

Earth System Grid

Climate change research is not only a scientific challenge of the first order – it is also a major technological challenge

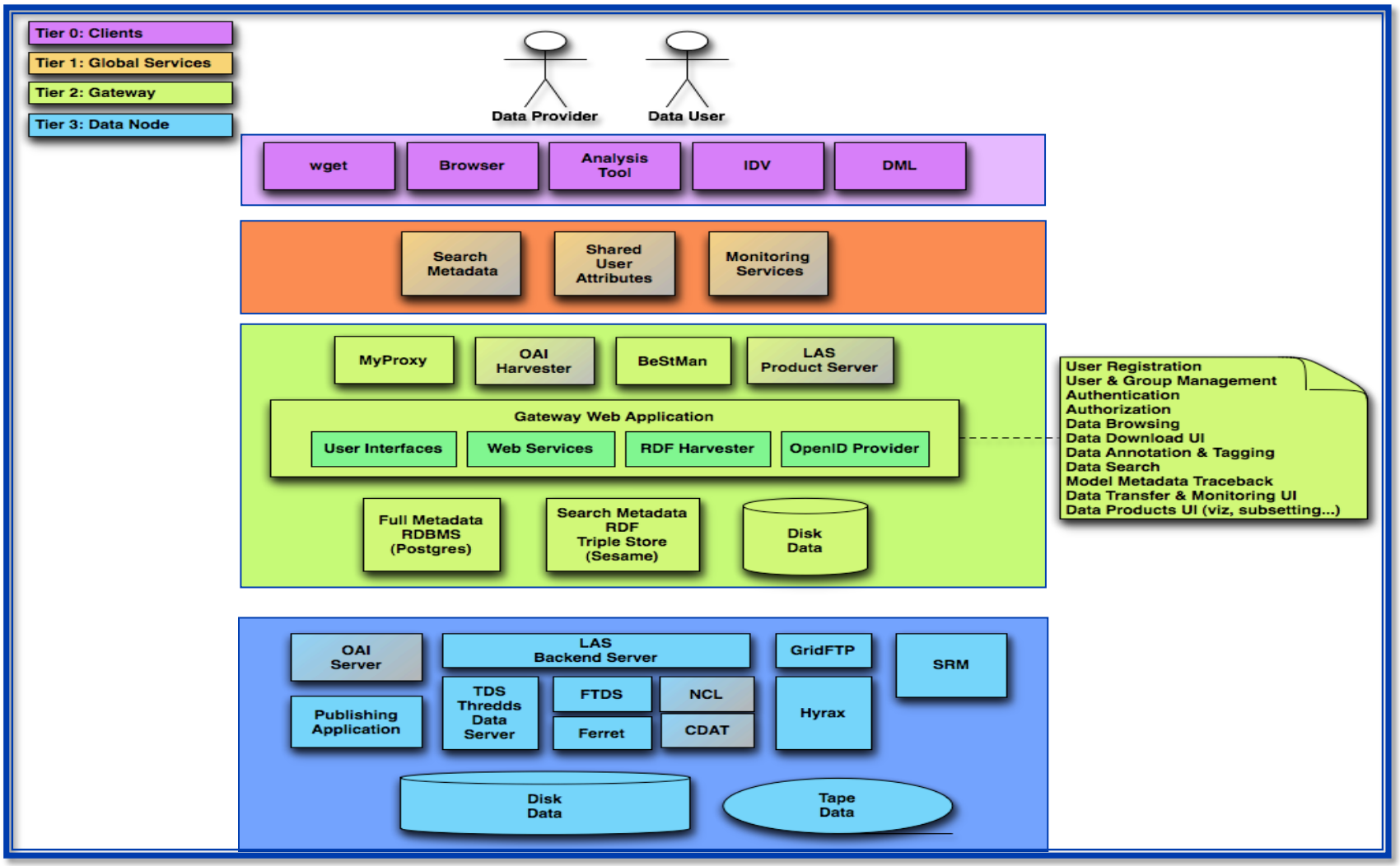


Technologies

the globus alliance · OPeNDAP · OpenID · LAS NOAA PMEL Live Access Server · Climate Data Analysis Tools (CDAT) · THREDDS · RDF · ESnet · SDM · OPEN ARCHIVES · Ferret NOAA PMEL · VACET · NCL NCAR Command Language

Architecture of Earth System Grid

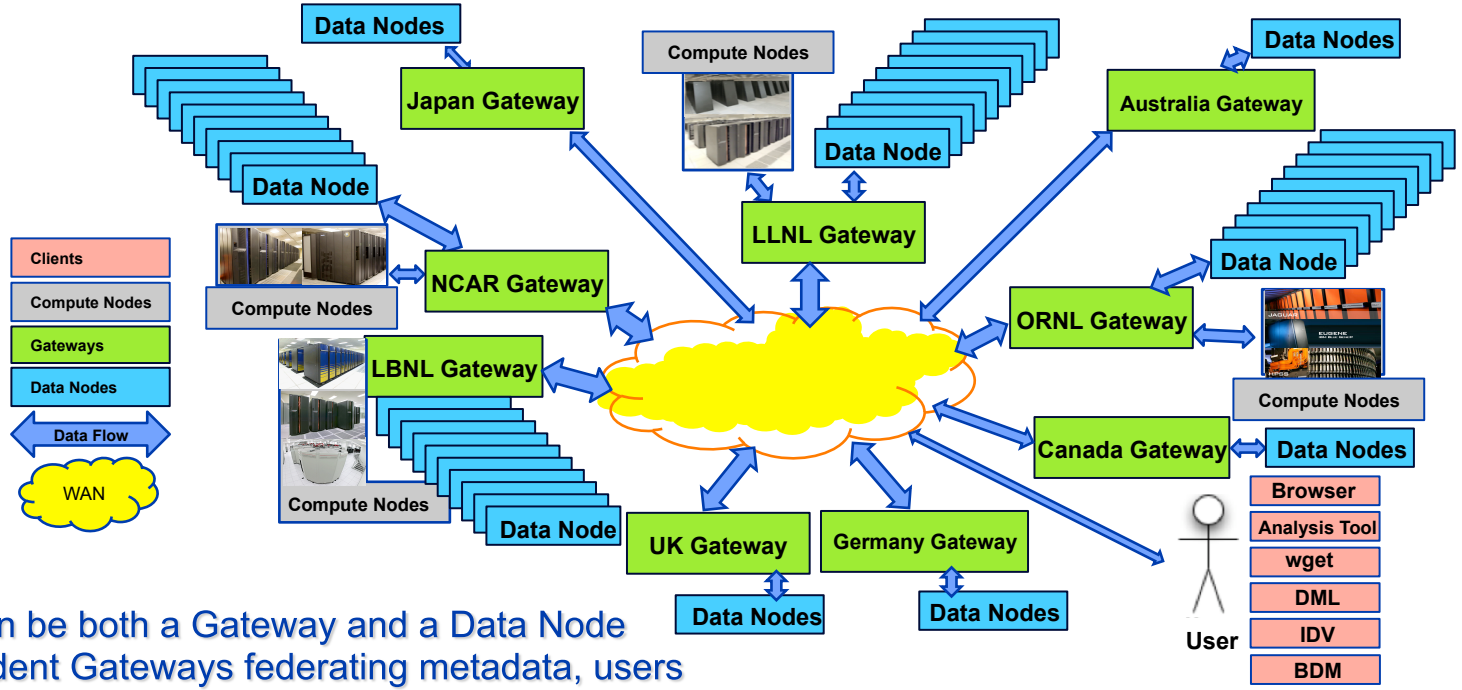
The ESG Gateway is a site which supports portal services.
End users interact with a portal to search and download data and data products.



Gateways and Data Nodes in ESG

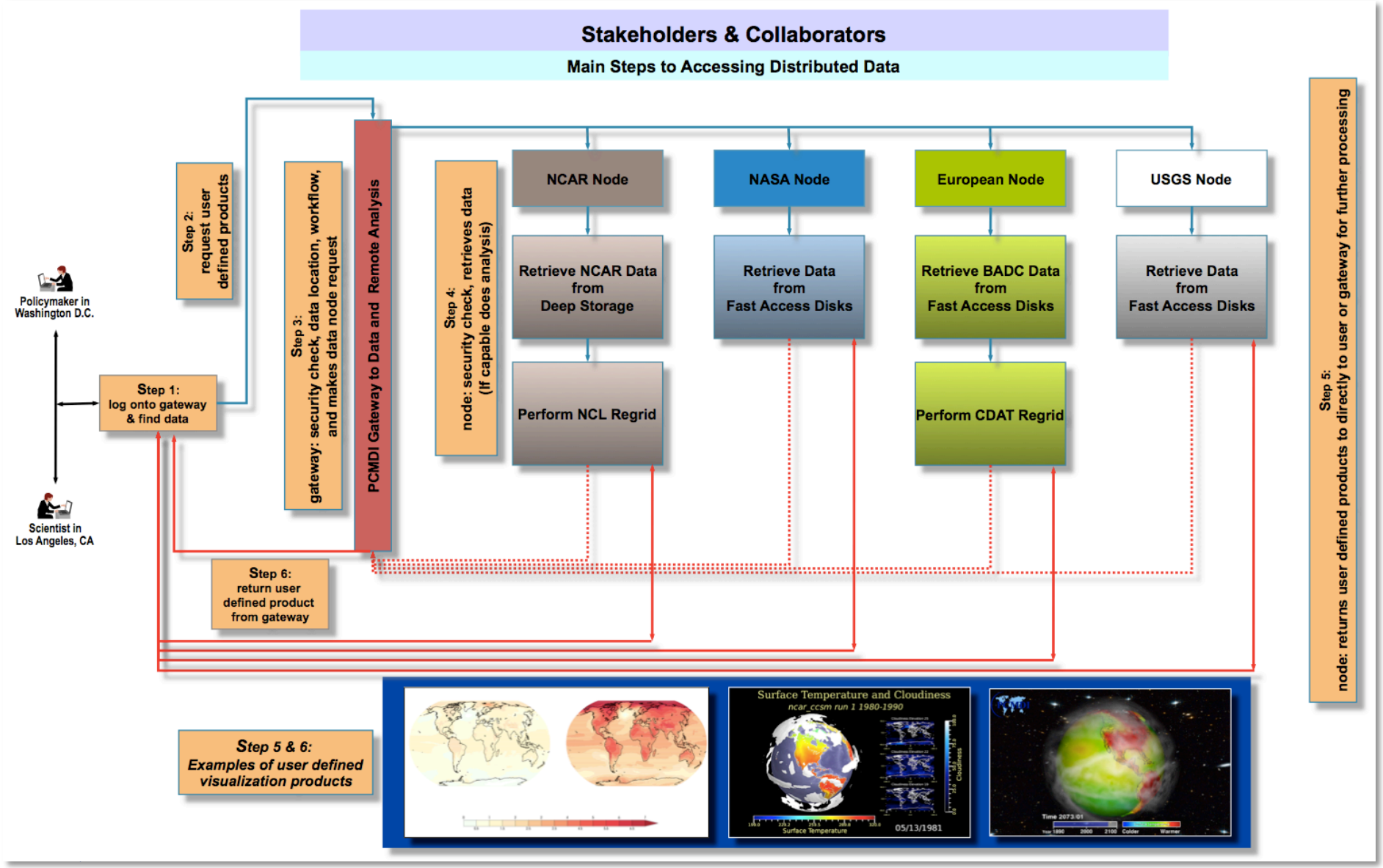
- Gateway Nodes: where data is discovered and requested
- Gateway Node functionality includes
  - portals, registration and user management
  - search capability, distributed metadata
  - may be customized to an institution's requirements, topical focus

- Data Nodes: where data is actually stored or archived
- Data Node functionality includes
  - data publish (making it visible to an ESG Gateway)
  - data reduction/analysis support
  - possible minimalist deployment without services
  - delivery services to ESG end users



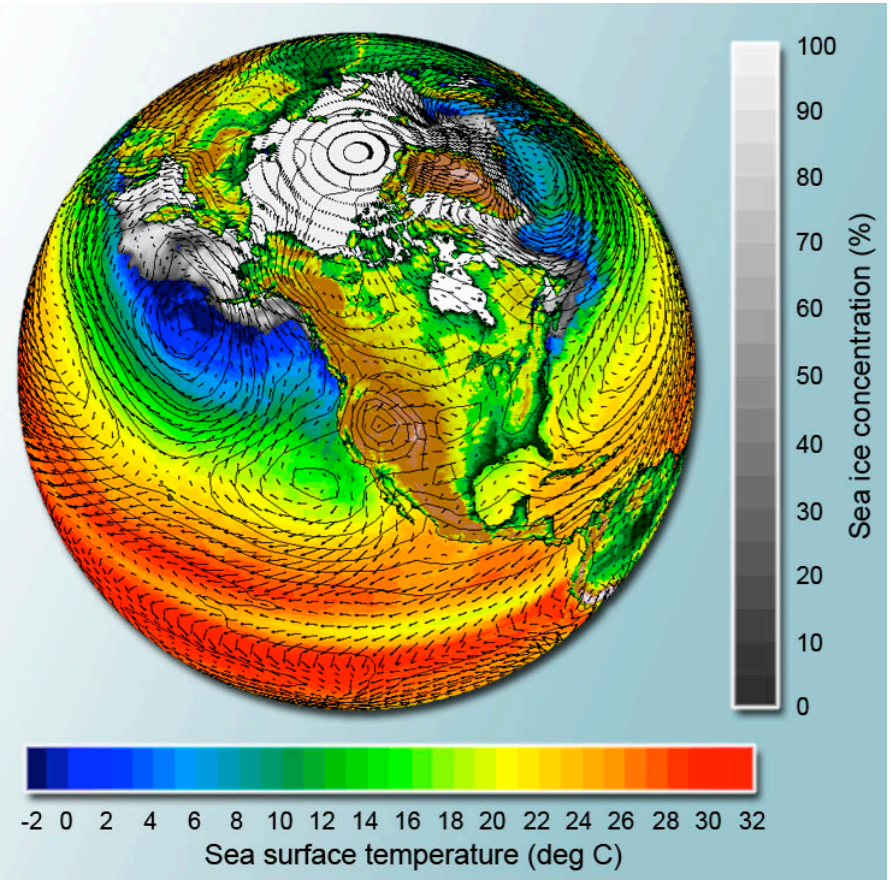
- A site can be both a Gateway and a Data Node
- Independent Gateways federating metadata, users
- Any user can discover any data from any Gateway
- Each data node publishes to one or more Gateways
- Specific data collections are managed through specific Gateways

ESG Use Case


Stakeholders & Collaborators
Main Steps to Accessing Distributed Data
Policymaker in Washington D.C.
Scientist in Los Angeles, CA
Step 1: log onto gateway & find data
Step 2: request user defined products
Step 3: gateway: security check, data location, workflow, and makes data node request
PCMDI Gateway to Data and Remote Analysis
Step 4: node: security check, retrieves data (If capable does analysis)
NCAR Node
NASA Node
European Node
USGS Node
Retrieve NCAR Data from Deep Storage
Retrieve Data from Fast Access Disks
Retrieve BADC Data from Fast Access Disks
Retrieve Data from Fast Access Disks
Perform NCL Regrid
Perform CDAT Regrid
Step 5: node: returns user defined products to directly to user or gateway for further processing
Step 6: return user defined product from gateway
Step 5 & 6: Examples of user defined visualization products
Surface Temperature and Cloudiness
ncar_ccsm run 1 1980-1990
05/13/1981

The Growing Importance of Climate Simulation Data

- **Broad investments in climate change research**
  - Development of climate models
  - Climate change simulation
  - Model intercomparisons
  - Observational programs
- **Climate change research is increasingly data-intensive**
  - Analysis and intercomparison of simulation and observations from many sources
  - Data used by model developers, impacts analysts, policymakers



Results from the Parallel Climate Model (PCM) depicting wind vectors, surface pressure, sea surface temperature, and sea ice concentration. Prepared from data published in the ESG using the FERRET analysis tool by Gary Strand, NCAR.

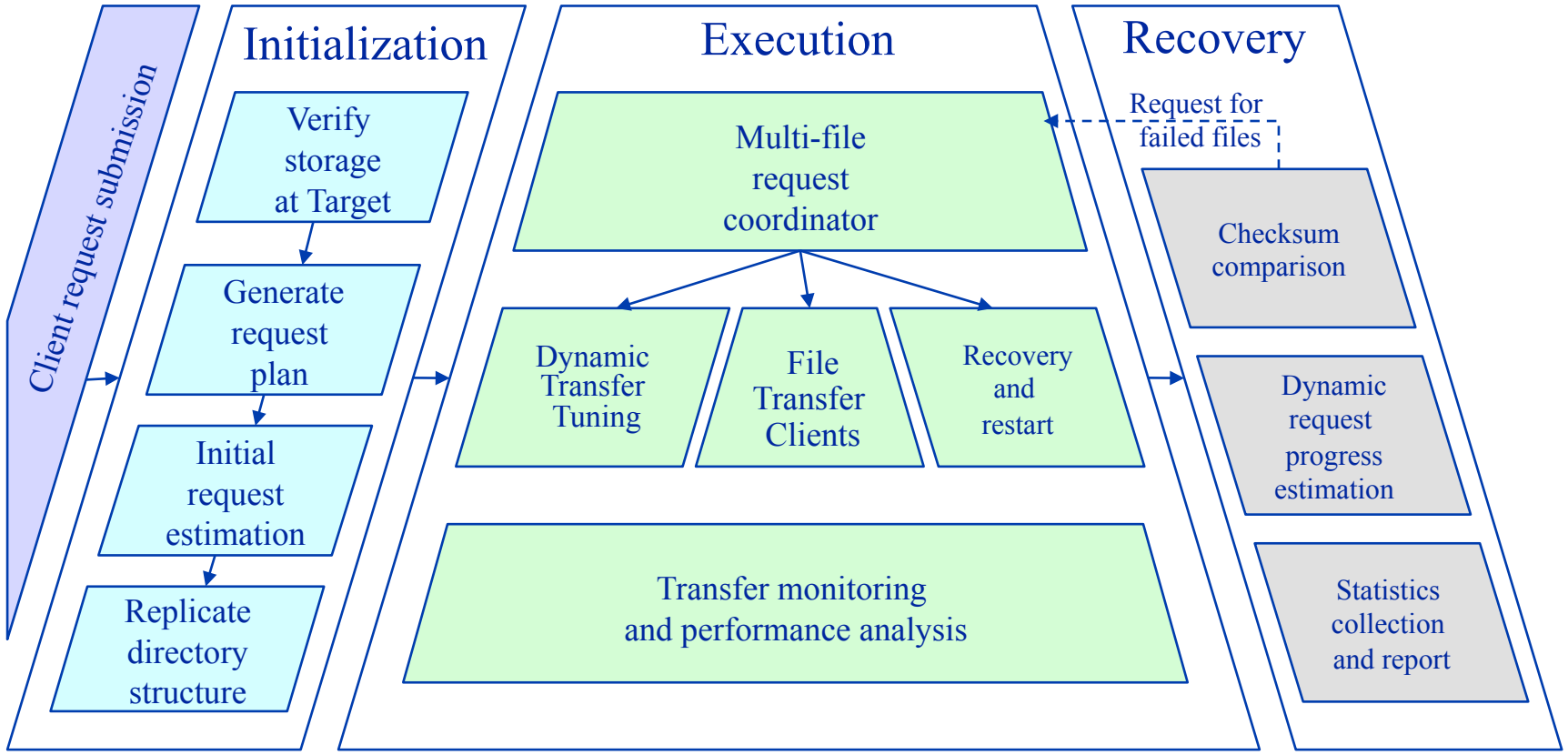
The Growing Size of Climate Simulation Data

- Early 1990's (e.g., AMIP1, PMIP, CMIP1)
  - modest collection of monthly mean 2D files: ~1 GB
- Late 1990's (e.g., AMIP2)
  - large collection of monthly mean and 6-hourly 2D and 3D fields: ~500 GB
- In 2000's (e.g., IPCC/CMIP3)
  - fairly comprehensive output from both ocean and atmospheric components; monthly, daily, and 3 hourly: ~35 TB
- In 2011:
  - The IPCC 5th Assessment Report (AR5) in 2011: expected 5 to 15 PB
  - The Climate Science Computational End Station (CCES) project at ORNL: expected around 3 PB
  - The North American Regional Climate Change Assessment Program (NARCCAP): expected around 1 PB
  - The Cloud Feedback Model Intercomparison Project (CFMIP) archives: expected to be .3 PB
- CMIP5 is being defined now, available info neither complete nor final
  - Current estimates… 1.2 to 2 PB of "replica core" results (35 TB for CMIP3)
    - In CMIP5, "replica core" is expected to be 20% t0 30% of total volume of data produced

Emerging Requirements

- Data search, accessibility, versioning, metrics, …
- Replication of core dataset
  - Disaster protection, support IPCC WG2 & WG3
  - Replication - A node can choose to replicate a collection of the datasets published by a different node. This includes replicas of aggregated datasets.
  - Bulk Data Movement - The transfer of large collections between sites reliably and with good performance and with a high-level of easy-of-use.
- Bulk Data Movement Requirements
  - Transfer in "pull mode" for security reasons
  - Scalability: Large in volume and number of files
    - Move terabytes to petabytes (many thousands of files)
    - Efficient handling on extreme variance in file sizes
  - Reliability and Robustness
    - Asynchronous long-lasting operation
    - Recovery from transient failures and automatic restart
    - Space verification at target storage
    - Support for checksum verification
  - On-demand transfer request status
    - Estimation of request completion time
    - Statistics collection
  - Multiple transfer protocol support
    - Use GridFTP and other transfer protocols if necessary
    - Take advantage of network provisioning

Architecture of Bulk Data Mover



Initialization phase in BDM

- Plans and prepares file replications from the data source to the target storage
  - Verification of target storage
    - Getting the total size of the request from the source site
    - Checking sufficient target storage allocation
  - Generating a request plan
    - Initial level of concurrency, number of parallel streams, and buffer size
    - Based on the historical statistics from the previous requests, if available
    - Otherwise, conservative default values could be used
    - These values could be adjusted based on the performance, as the request progresses.
  - Initial request estimation based on the request plan
  - Generates at the target site a mirror image of the directory structure at the source.
  - Generates an execution plan
    - includes pair-wise source-to-target URLs for all the files to be replicated.
    - This is used by the execution phase.

Execution Phase in BDM

- **Transfers the requested files, while monitoring and analyzing transfer performance**
  - **Multi-file request coordinator**
    - uses the information from the execution plan
    - instantiates multiple instances of the file transfer client
  - **File transfer client**
    - supports any transfer protocols or services preferred by the community
    - supported transfer protocols could be GridFTP, HTTPS, SCP, SFTP, etc.
  - **Recovery and restart module**
    - continuously monitors the health of the system and the files being transferred
    - If a transient error occurs, reschedules the transfer or continues the transfers from the point of interruption
  - **Monitoring and analysis module**
    - monitors and collects dynamic transfer performance
    - if discrepancies from the estimated performance are noticed, it adjusts the number of concurrency and parallel streams

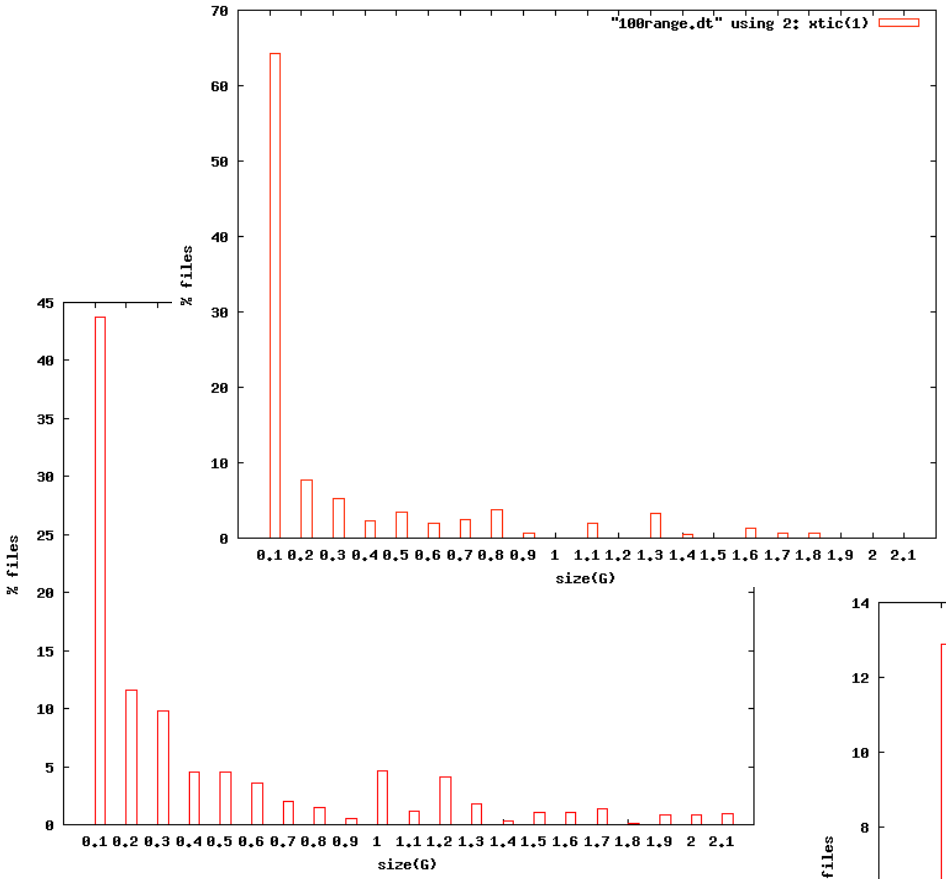
Recovery Phase in BDM

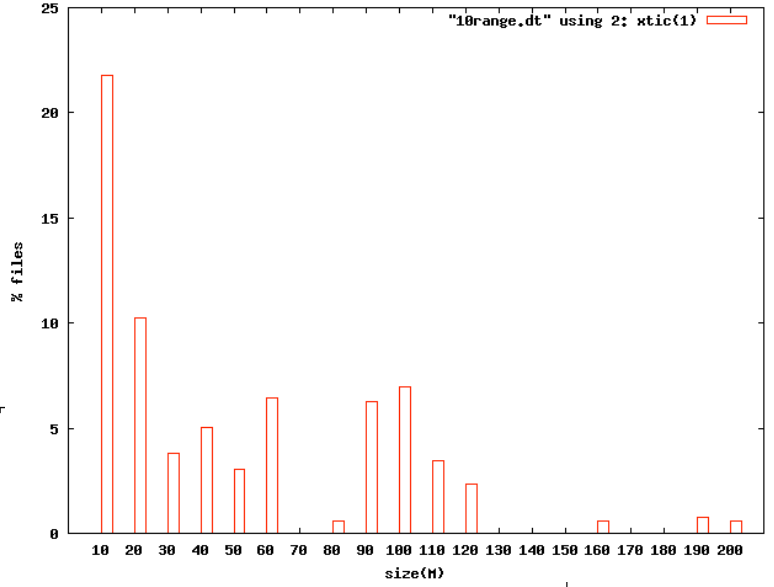
- **Dynamic interaction with execution components**
  - validates the completed transfer request
  - collects statistics from the execution of the replication request
  - dynamic progress estimation on-demand
  - file validation by checksum comparison
  - re-submitting file transfers whose checksums indicated data corruption

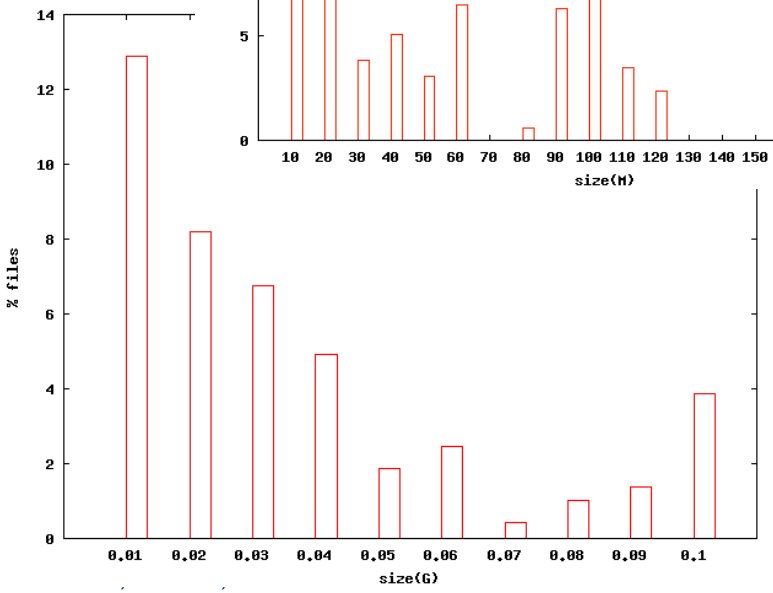
Transfer Performance in BDM

- **High performance using a variety of techniques**
  - Multi-threaded concurrent transfer connection management
  - Transfer queue management
  - Single control channel management for multiple data transfers
  - Load balancing on multiple transfer servers
  - GridFTP library supports data channel caching and pipelining
- **Transfer queue and concurrency management**
  - Contribute to more transfer throughput, including both network and storage
  - When a dataset contains many small files, concurrent transfers with overlapping storage I/O with the network I/O helps improve the transfer performance

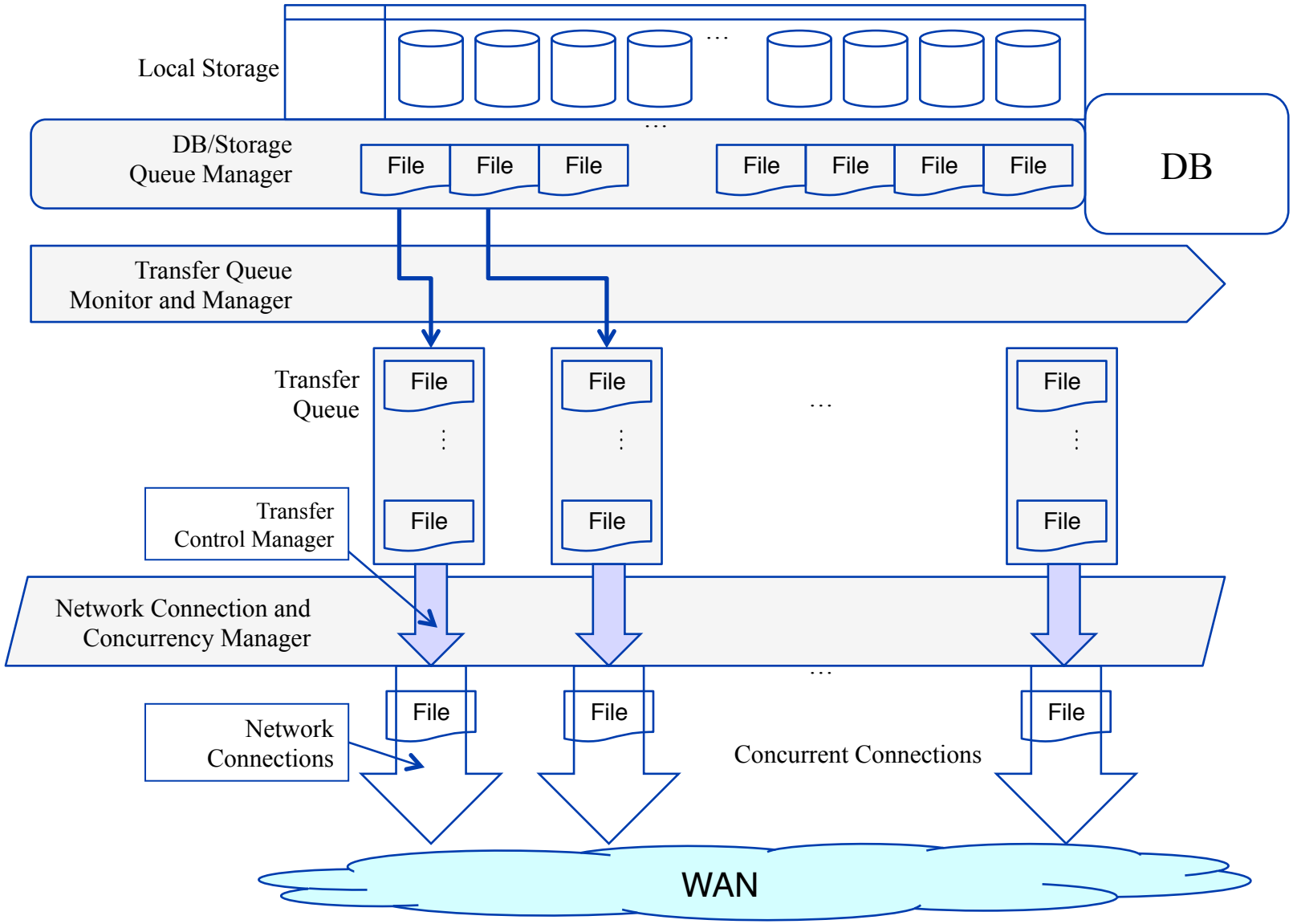
Distribution of Files – Characteristics of datasets







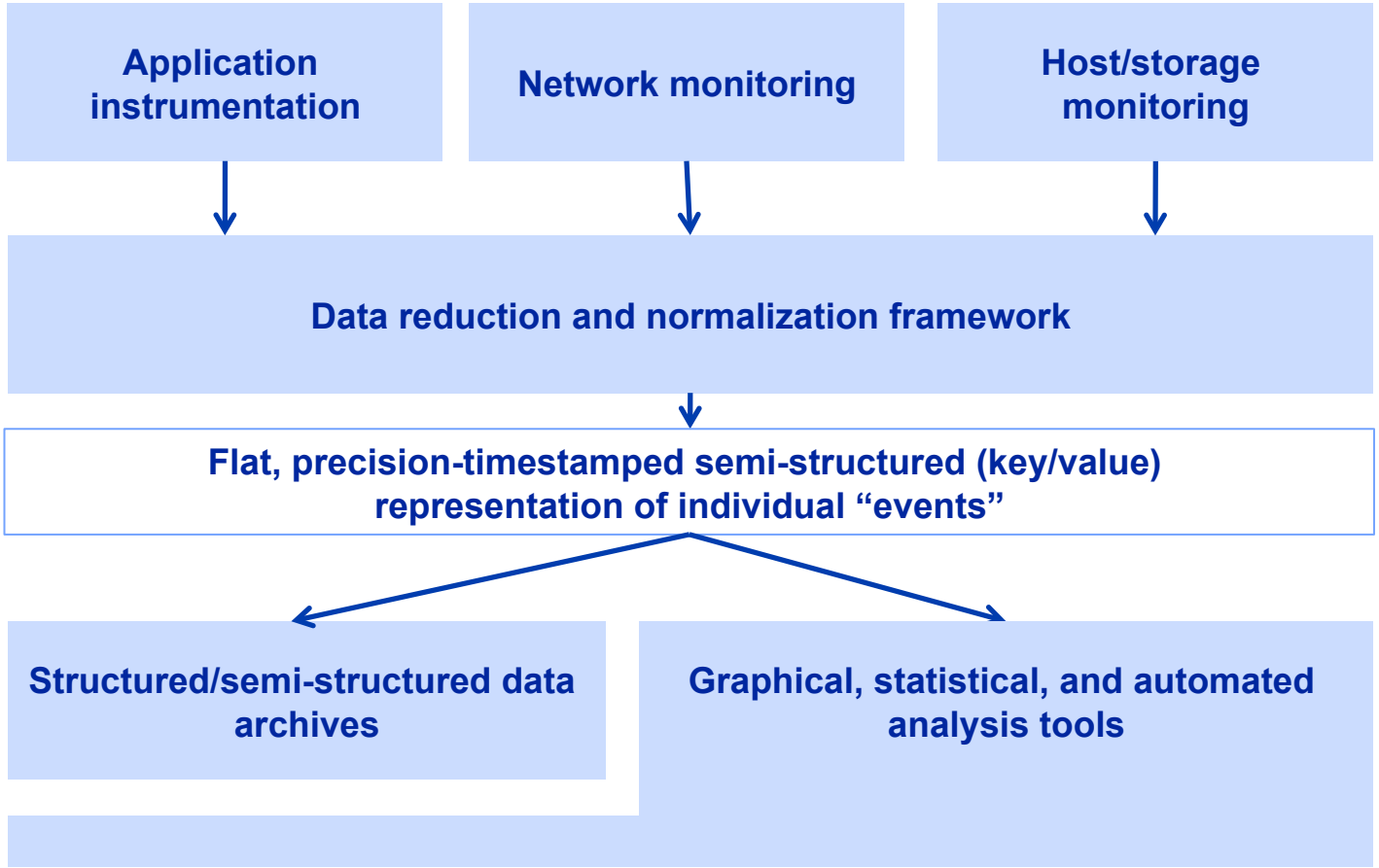
Transfer Queue and Concurrency Management in BDM



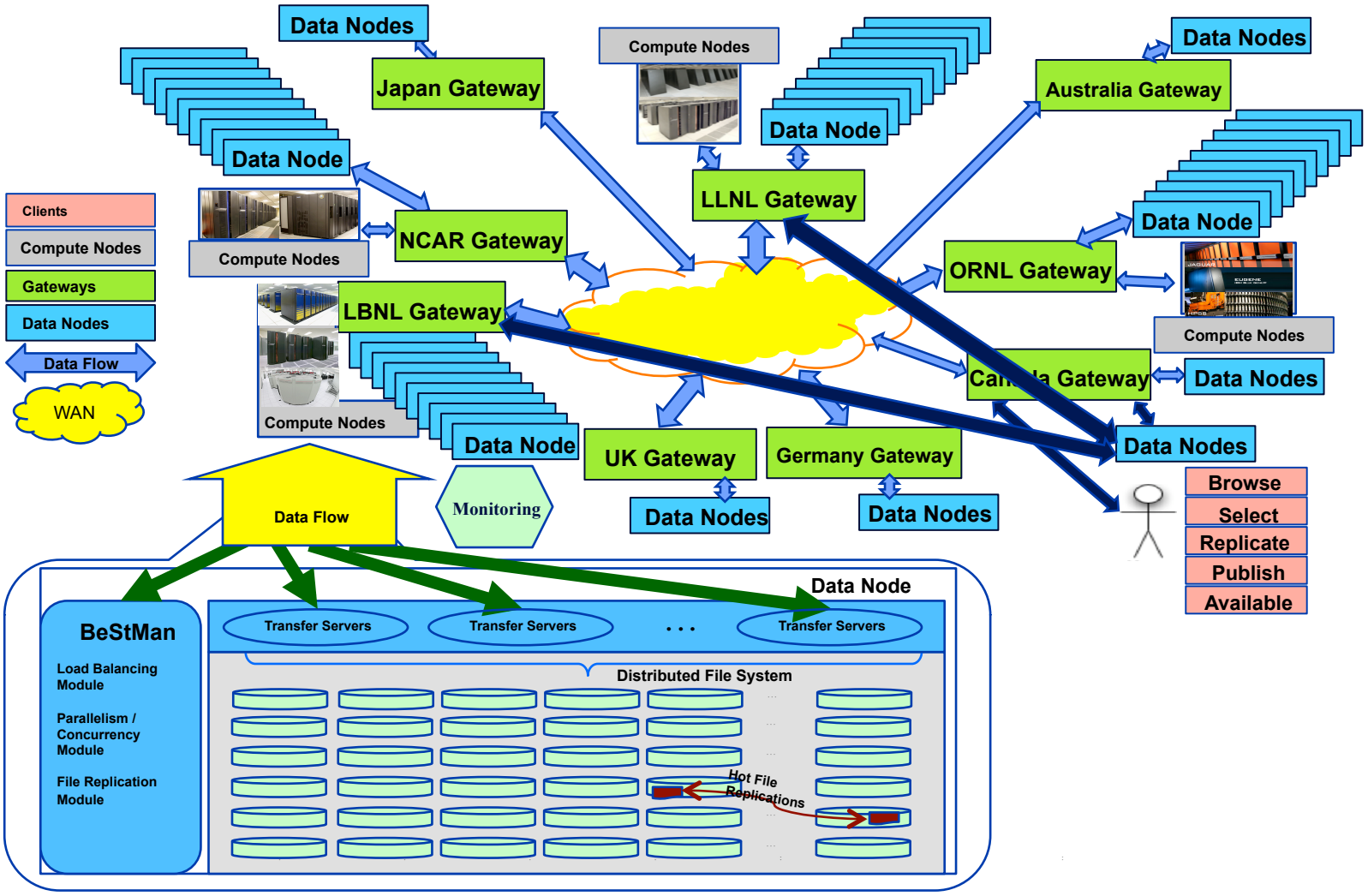
NetLogger Monitoring Overview

- **Goals**
  - End-to-end distributed performance analysis and troubleshooting
- **Focus areas**
  - Wide-area data transfers
  - Distributed workflows
  - Service-oriented computing
- **Complementary technologies**
  - *syslog-ng* log collection software
  - *R* statistical data analysis environment

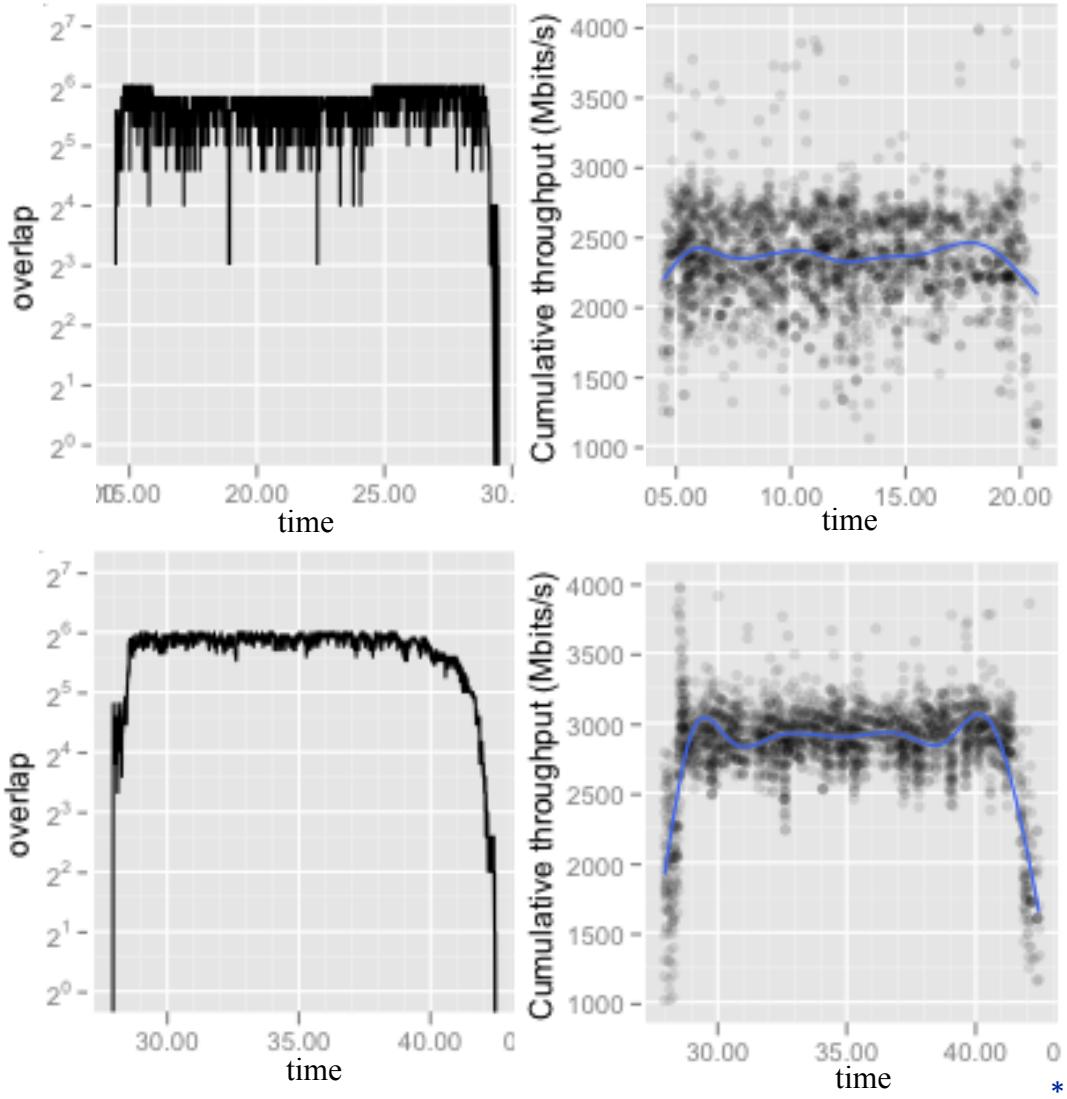
NetLogger components



Replication Use Case



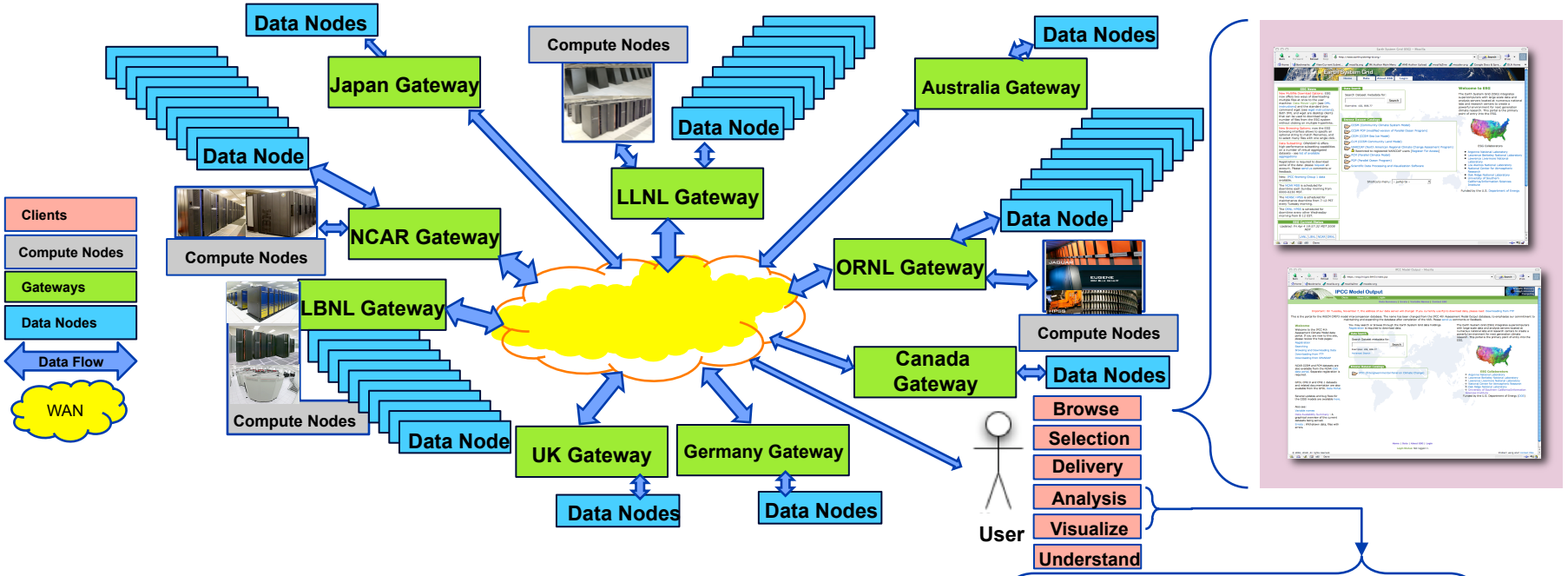
Results of Managed Transfers

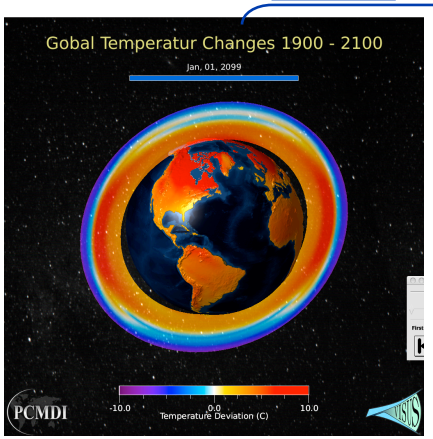


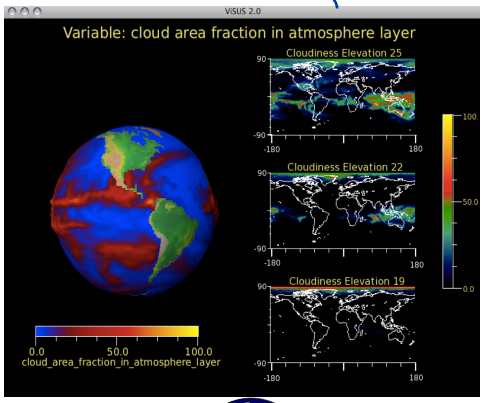
The number of concurrent transfers on the left column shows consistent over time in well-managed transfers shown at the bottom row, compared to the ill or non-managed data connections shown at the top row. It leads to the higher overall throughput performance on the lower-right column.

* Plots generated from NetLogger

All together…







- **Enable new capabilities for analysis of data and visual exploration**
  - Visualization of uncertainty and ensemble data
- **Help scientists understand long-term climate impact**

Information

- **Bulk Data Mover**
  - http://sdm.lbl.gov/bdm/
- **NetLogger Monitoring**
  - http://dsd.lbl.gov/NetLoggerWiki/
- **Earth System Grid**
  - http://www.earthsystemgrid.org
  - http://esg-pcmdi.llnl.gov/
- **Support emails**
  - esg-support@earthsystemgrid.org
  - srm@lbl.gov